Phrasal constructs, the lexicon, and multilingual generation

The lexicon is normally conceived of as the repository of word-specific information. Traditional lexical resources, such as machine readable dictionaries, therefore contain lists of words. These lists might delineate senses of a word, represent the meaning of a word, or specify the syntactic frames in which a word can appear, but the level of granularity with which they are concerned is the individual word. There are many linguistic phenomena which pose a challenge to this "word focus" in the lexicon. The incorporation of elements at a higher level of abstraction — at the phrasal level, where particular words are grouped together into fixed phrases — provides a basis for improved computational processing of language. In this talk, I will examine the phrasal lexicon and discuss its utility in the context of multilingual text generation.

It comes as no surprise to any student of language that text analysis at the level of individual words ignores many important components of language (Becker 1975, *inter alia*). To list but a few examples, languages contain idioms (e.g. *to kick the bucket*), fixed expressions (e.g. *by and large*, *to boldly go*), phrasal proper nouns (e.g. *The Big Apple*), collocations (e.g. *eye dropper*, *red herring*), and phrasal verbs (e.g. *to go out*, *to look up*). In general, such phrasal constructs cannot be interpreted on the basis of a compositional analysis of their constituent words, but instead have specific interpretations which may or may not be related in some way to the meanings of their constituent words and which demand a lexical representation independent of those constituents.

Another type of phrasal construct abstracts away from individual words entirely: particular syntactic configurations can be associated with a specific interpretation. Goldberg (1995), for example, argues that constructions such as ditransitives ([NP_{subj} V NP_{obj1} NP_{obj2}]: *Pat faxed Bill the letter*) are instances of fixed form-meaning correspondences which have syntactic and semantic characteristics not entirely derivable from composition of the individual words.

It therefore must come as no surprise that the lexicon must contain phrasal elements in order to provide an adequate resource for natural language processing. In the absence of such phrasal elements, the lexicon is an inadequate resource for the interpretation of natural language texts and for addressing the conventional aspects of language use.

Apart from the theoretical need to address phrasal constructs in language, there are computational gains which stem from using a phrasal lexicon. Where a concept is expressed as a phrase rather than as an individual word, a system which utilises a phrasal lexicon can avoid breaking that phrase down into its constituents and attempting to construct a (non-existant or inaccurate) compositional interpretation for it. The phrase can simply be treated as a single coherent constituent, with an associated meaning and without internal structure. A system which encounters such a phrase need not spend time or resources to analyse it further.

The use of a phrasal lexicon has been particularly successful in the area of natural language generation, where a system normally has a finite set of complex concepts which can be expressed in language. Milosavljevic, Tulloch and Dale (1996), for example, describe the PEBA-II system, which utilises a phrasal lexicon. The system has a knowledge base in which not only simple entities and properties are represented, but also higher-level semantic objects. These objects correspond to complex constructions which make "precisely those distinctions that are relevant for the range of texts [they] intend to generate". Each of these conceptual abstractions is associated with a phrasal realisation in the lexicon. In their domain of animal descriptions, for example, making explicit comparisons between animals is important, but only certain categories of comparisons are required. A complex statement such as *The Shrew has soft fur* will be represented as `(has-prop Shrew (body-covering soft-fur))` because it is important to be able to compare one animal's body covering with another's (e.g. *The Shrew has soft fur while the Mole has thick velvety fur*). It is not, however, necessary to break the kind of body covering, `soft-fur`, down into smaller representational elements (e.g. `(fur`

`(has-prop soft))`, as more complex comparisons (e.g. *The Shrew and the Mole both have fur, but the Shrew's is soft while the Mole's is thick and velvety*) are not relevant for the target texts. The concept `soft-fur` is then directly associated with the realisation "has soft fur" in the lexicon.

This approach avoids both the need for generation of syntactically complex realisations from individual words and expensive construction of a knowledge base which is more complex than necessary for the domain. The level of abstraction in the knowledge representation is mirrored in the level of decomposition in the lexicon, and the system can more efficiently generate texts expressing particular concepts. It is better to explicitly represent the realisation of certain complex concepts which are repeatedly realised the same way, to avoid rebuilding the surface form for each occurrence (Milosavljevic, Tulloch and Dale 1996).

The use of a phrasal lexicon clearly has benefits for monolingual generation systems for which the level of granularity of elements in the knowledge representation can directly correspond to the level of granularity in the lexicon. We are currently exploring the extension of these techniques to multilingual text generation — that is, generation of texts in several different languages on the basis of a language-independent underlying representation and language-specific phrasal lexica and grammars. The approach is similar to that of example-based machine translation, in that concepts are not always broken down into atomic elements, but it avoids some of the difficulties inherent in machine translation systems, takes advantage of the use of underlying structured data, and builds in the potential for flexibility and dynamism in the output text (Hartley and Paris 1997).

The decomposition of information for natural language generation must meet two competing desiderata:

- **Coverage:** Representational elements must be fine-grained enough to capture the full range of concepts and predicates which are relevant to the domain.
- **Reusability:** Representational elements must be broad enough to reflect generalisations which can be made about related concepts.

The difficulties inherent in balancing these desiderata are amplified in the multilingual context, because different languages place competing constraints on the granularity of the "language-independent" knowledge representation. An individual concept may correspond to hugely different surface realisations in different languages, and the appropriate level of granularity for the representation of concepts may differ for different languages. Further difficulties for multilingual text generation based on a phrasal lexicon stem from the fact that techniques for automatic construction of a knowledge base and target-language phrasal lexica are difficult to envisage. Construction of these components by hand is arduous and time-consuming.

The use of a phrasal lexicon can harness systematicity in the realisation of certain concepts in a language for efficiency gains in generation. We will report on our attempts to establish appropriate cross-linguistic generalisations which build on these techniques.

References

Becker, J. D. (1975). The phrasal lexicon. In *Proceedings of the Conference on Theoretical Issues in Natural Language Processing*, Cambridge, MA, pp. 70–77.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. The University of Chicago Press.

Hartley, A. and C. Paris (1997). Multilingual document production: From support for translating to support for authoring. *Machine Translation 12*(1-2), 109–129.

Milosavljevic, M., A. Tulloch, and R. Dale (1996). Text generation in a dynamic hypertext environment. In *Proceedings of the 19th Australasian Computer Science Conference*, Melbourne, Australia.